**Hierarchical clustering** (aka hierarchical cluster analyses, HCA)

**What is it?**

HCA is a method of cluster analysis, based on building hierarchies of clusters. There are two approaches:

1. *Agglomorative (bottom-up)*: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. (“moving your threshold”)
2. *Divisive (top-down)*: all observations start in one cluster, and groups are split as one moved down the hierarchy. (“changing your threshold”)

Generally, the clusters (or splits) are determined based on a greedy algorithm (or heuristic approach). In simplest way: at each step, or here for each observation, find the closest look-a-like. This does not necessarily lead to the best solution, but it is an efficient strategy.

Finding the most optimal solution for a HCA question, typically would require an unreasonable number of steps (and thus computation power).

A solution to this problem would be to run a lot of individual greedy-algorithm based hierarchical cluster analyses and add a value to the one found most regularly (e.g. a bootstrap based approach).

**Complexity and solutions**

As said, HCA’s are complex to run with a time complexity of (n3) and memory requirement of (n2). There are solutions to this complexity (that only help/work for specific cases):

* Single-linkage clustering (bottom-up) -> single-linkage is the shortest distance between **a pair** of observations in two clusters (nearest neighbour). It can happen here that observations in different clusters are closer than two observations in a single cluster.
* Complete-linkage clustering (top-down) -> at each step, two clusters separated by the shortest distance are combined (farthest neighbour)

The results of a hierarchical clustering are usually presented in a dendrogram, a diagram showing the hierarchical relationships between objects.

**Application**

Hierarchical clustering is used in phylogenetic relationship reconstructions/analyses. A phylogenetic analysis analyses the evolutionary relationship between species.

E.g. *Homo sapiens* (human)is closest related to *Pan paniscus* (bonobo), and both are closely related to *Pan troglodytes* (chimpansee).

These evolutionary relationships are based on morphological characteristics, or nowadays DNA data. The relationships are “basically” based on hierarchical clustering methods.